

## Informe Final projecte 2017PID-UB/034

### Biblioteca d'aplicacions Shiny per a un primer curs d'Anàlisi de Dades Multivariants

**Responsable:** Ferran Reverter Comes

**Departament:** Genètica, Microbiologia i Estadística

**Facultat:** Facultat de Biologia

**Correu electrònic:** freverter@ub.edu

## RESUM I DESCRIPTORS

### Resum

El projecte suposa la continuïtat del projecte Biblioteca d'aplicacions Shiny per a un primer curs d'Anàlisi de Dades Multivariants, acollit en la segona convocatòria d'ajuts Universitat de Barcelona-Banco de Santander en el marc del programa de retenció de talent per mèrits docents (2016).

L'actual projecte es plantejava **ampliar** la biblioteca d'aplicacions Shiny per tal d'ajudar a comprendre algunes de les tècniques fonamentals d'Anàlisi de Dades Multivariants: Anàlisi de Components Principals - Biplot, Anàlisi Clúster, k-means i Classificadors (Lineal, Quadratic, K-Nearest Neighbor, Màquines de Vector de Suport). En el projecte, a més a més, es persegueix la implementació d'una **primera ronda de pràctiques** amb estudiants per recollir el seu *feedback* general sobre l'aplicatiu.

Cada una de les aplicacions que constitueixen la biblioteca Shiny mostren un recorregut guiat per les qüestions essencials de cada tècnica això, creiem, facilitarà un apropament ordenat als aspectes rellevants.

D'acord amb els plans d'estudis dels graus de Bioquímica, Biotecnologia i Ciències Biomèdiques de la Facultat de Biologia, inclouen l'assignatura de Disseny d'Experiments i Anàlisi de Dades a quart curs amb plans docents equivalents que defineix el context d'aplicació del projecte. El Disseny

d'Experiments i Anàlisi de Dades (6 ECTS) és una assignatura obligatòria de quart curs on es presenten els conceptes bàsics del disseny experimental i les tècniques fonamentals d'Anàlisi de Dades multivariants. Aproximadament els dos blocs de l'assignatura es reparteixen en un 80%-20%, respectivament.

Dels resultats previstos, s'ha assolit el primer i no el segon. S'ha desenvolupat una nova activitat per ampliar la biblioteca però no hem arribat a temps per fer la ronda d'interacció amb els estudiants.

## Descriptors

- Línies d'innovació vinculades

- Simulació,
- Autoavaluació
- Aprenentatge autònom

- Paraules clau

- Aplicació Shiny
- Anàlisi Multivariant
- PCA
- K-means
- SVM
- Clustering

## MANCANCES DETECTADES

El context d'aplicació correspon a l'assignatura de Disseny d'Experiments i Anàlisi de Dades (DEAD) en els graus de Bioquímica, Biomedicina, Biotecnologia. Els estudiants sovint mostren dificultats per copsar les nocions **geomètriques i estadístiques** subjacents en les tècniques multivariants. Va semblar oportú implementar aquesta biblioteca d'aplicacions basades en Shiny, que a partir de la interacció guiada per l'aplicatiu faciliti l'autoaprenentatge dels estudiants.

## OBJECTIUS

- 1) Ampliar la biblioteca d'aplicacions Shiny per donar cabuda a altres tècniques d'anàlisi de dades d'interès pels estudiants.

2) Implementar les activitats fent èmfasi en els aspectes interactius i visuals.

3) Posar en pràctica en sessions reals l'ús de la biblioteca d'aplicacions per a un primer curs d'anàlisi de dades multivariants.

## **DESENVOLUPAMENT DE L'ACTUACIÓ**

S'ha implementat una activitat nova per a la docència dels classificadors de vector de suport (SVM) que s'afegeix a la biblioteca d'aplicacions desenvolupades anteriorment. Els classificadors SVM son una de les metodologies més potents de l'aprenentatge màquina prèvies a l'eclosió del Deep Learning. Tenen una sòlida fonamentació estadística i a la vegada una intuïtiva interpretació geomètrica basada en la noció de marge màxim entre classes, cosa que fa dels SVMs un tòpic idoni per introduir als estudiants noves metodologies per a la classificació/regressió que milloren a les clàssiques. En l'apèndix es poden trobar detalls de la implementació de l'activitat.

Cal comentar que les quatre aplicacions que constitueixen la biblioteca Shiny son una eina per complementar els dos elements en que es recolzen les sessions habituals de l'assignatura de Disseny d'Experiments i Anàlisi de Dades: d'una banda, l'exposició dels conceptes estadístics i de les eines computacionals, i d'altra banda, l'anàlisi i la interpretació de resultats a partir de casos pràctics. El material desenvolupat és un nou element, es pot dir que cau en mig dels dos anteriors, no es teoria *per se* però tampoc anàlisi de dades. Una vegada el professor ha introduït i explicat els principis i característiques de la tècnica multivariant en qüestió (slides, pissarra, ...) i abans d'entrar en l'anàlisi de casos, es pot indicar als estudiants l'ús del material per reafirmar l'aprenentatge dels conceptes propis de cada tècnica.

## AVALUACIÓ, RESULTATS I INTERPRETACIÓ

La valoració de l'actuació que s'ha portat a terme en aquest projecte d'innovació docent ha sigut positiva. S'ha aconseguit ampliar el número d'activitats disponibles en la biblioteca que ara inclou els classificadors SVM.

Ha mancat la ronda amb els estudiants per recollir el seu *feedback*, però pal·liar fins a cert punt aquesta mancança, s'han fet sessions amb professors del Departament per mostrar-los l'aplicatiu i s'han recopilat les suggerències per actualitzar i millorar els aspectes didàctics de les activitats desenvolupades.

## REFERÈNCIES BIBLIOGRÀFIQUES

Beeley, C. (2013). *Web application development with R using Shiny*. Packt Publishing Ltd.

<https://cran.r-project.org/>

Ripley, B. D. (2001). The R project in statistical computing. *MSOR Connections. The newsletter of the LTSN Maths, Stats & OR Network*, 1(1), 23-25.

Peña, Daniel. *Análisis de datos multivariantes*. McGraw-Hill España, 2013.

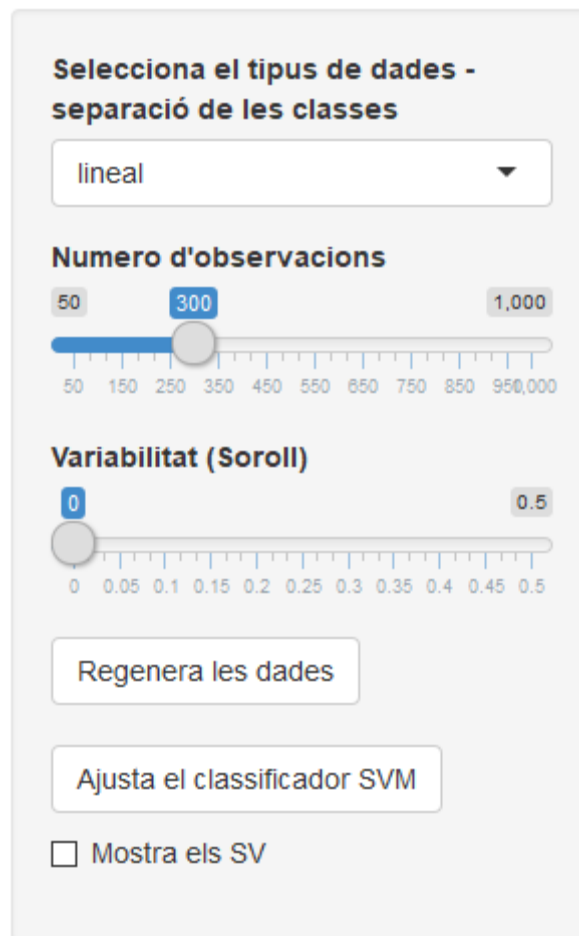
[http://www.ia.uned.es/~ejcarmona/publicaciones/\[2013-Carmona\]%20SVM.pdf](http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf)

<https://dlegorreta.wordpress.com/2015/04/07/maquina-de-soporte-vectorial-svm-soport-vector-machine/>

## Apèndix

L'aplicació associada a l'activitat SVM consta de dues parts, el sidebarPanel i el mainPanel. El sidebarPanel és on es situaran les opcions que l'estudiant haurà d'escollir i el mainPanel és on tindrem els gràfics segons la configuració del sidebarPanel.

El sidebarPanel és de la forma:



The sidebar panel is a light gray rectangular box containing the following elements from top to bottom:

- Selecciona el tipus de dades - separació de les classes:** A dropdown menu with the text "lineal" and a downward arrow.
- Numero d'observacions:** A horizontal slider. The track has major ticks at 50, 150, 250, 350, 450, 550, 650, 750, 850, 950, and 1,000. A blue slider knob is positioned at 300.
- Variabilitat (Soroll):** A horizontal slider. The track has major ticks every 0.05, from 0 to 0.5. A blue slider knob is positioned at 0.
- Regenera les dades:** A rectangular button.
- Ajusta el classificador SVM:** A rectangular button.
- Mostra els SV:** A checkbox followed by the text "Mostra els SV".

El primer element del sidebarPanel és una llista desplegable. A través de la que l'estudiant selecciona el tipus de separació que hi ha entre les classes que determinen el problema de classificació.

**Selecciona el tipus de dades -  
separació de les classes**

lineal

- lineal
- quadratic
- sinusoidal

Això permet comparar el rendiment del SVM davant de problemes que son més difícils de discriminar. Un segon element permet seleccionar el número d'observacions en cada classe, s'ha plantejat una situació balancejada.

**Numero d'observacions**

50 300 1,000

50 150 250 350 450 550 650 750 850 950 1,000

A continuació s'incorpora una element que permet introduir observacions que incompleixen el model de separació. Aquest paràmetre fa que un percentatge aleatori d'observacions del data set intercanviïn la classe i d'aquesta manera incrementa la dificultat de la tasca de classificació.

**Variabilitat (Soroll)**

0 0.5

0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5

La resta d'elements permeten generar el data set d'observacions i ajustar pròpiament el model SVM d'acord a les dades generades (cal indicar que en aquesta primera versió només s'ha implementat el kernel lineal, deixant per a noves versions la possibilitat d'altres kernels, com per exemple, el gaussià, polinomial, entre altres). Addicionalment, amb l'opció "Mostrar els SV" es mostren les observacions que son els vectors de suport.

Regenera les dades

Ajusta el classificador SVM

☐ Mostra els SV

A continuació es mostren dues imatges del MainPanel. La primera on s'aprecien les dades generades, indicant les observacions de cada una de les dues classes amb color blau i salmó. La segona mostra les dues regions discriminades pel model SVM. L'estudiant pot apreciar la qualitat del classificador.

## SVM

### Màquines de Vector de Suport (Support Vector Machines)

Per mostrar el classificador SVM, generarem el conjunt de dades de prova per a la classificació. Podem triar el tipus de separació de les dues classes: lineal, quadràtica o sinusoidal. Després ajustem el classificador SVM mostrem els resultats.

**Per iniciar - triar el tipus de separació, regenera les dades, i clicar Ajusta el classificador SVM.**

**Selecciona el tipus de dades - separació de les classes**  

sinusoidal

**Numero d'observacions**  

503001,000

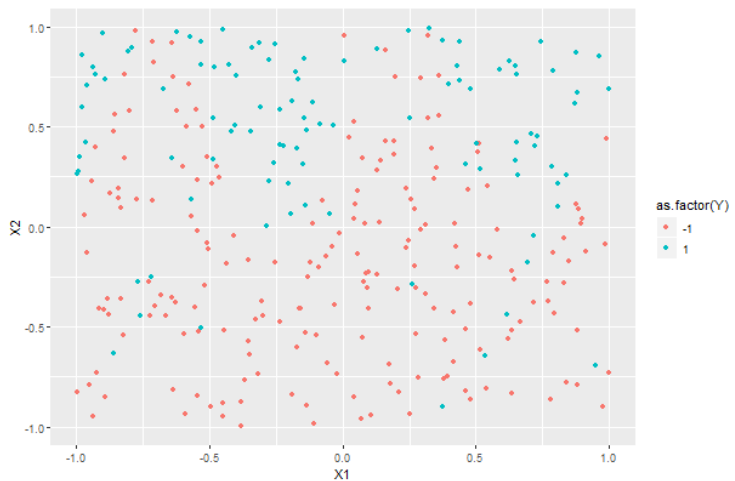
**Variabilitat (Soroll)**  

00.080.5

Regenera les dades

Ajusta el classificador SVM

☐ Mostra els SV

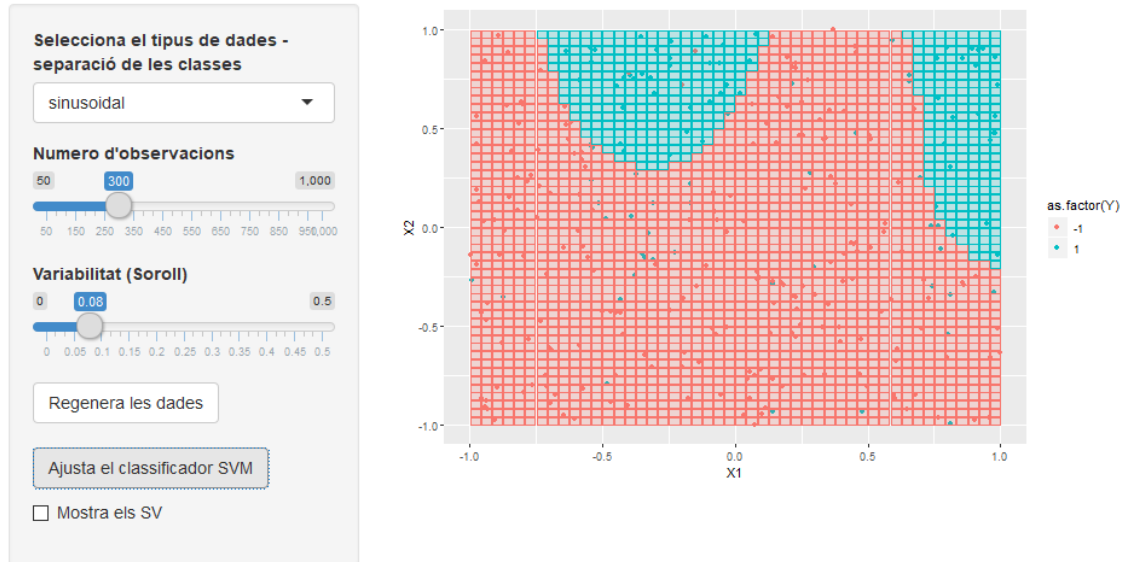


# SVM

## Màquines de Vector de Suport (Support Vector Machines)

Per mostrar el classificador SVM, generarem el conjunt de dades de prova per a la classificació. Podem triar el tipus de separació de les dues classes: lineal, quadràtica o sinusoidal. Després ajustem el classificador SVM mostrem els resultats.

**Per iniciar - triar el tipus de separació, regenera les dades, i clicar Ajusta el classificador SVM.**



Es deixa com exercici pels estudiants el càlcul de mètriques del classificador. No obstant, en properes versions s'inclouran en l'activitat noves pestanyes adreçades a les mètriques i a la validació creuada.